Umanistica Digitale - ISSN:2532-8816 - n.5, 2019
R. Sprugnoli, G. Pardelli, F. Boschetti, R. Del Gratta – Un'Analisi Multidimensionale della Ricerca Italiana nel Campo delle Digital Humanities e della Linguistica Computazionale
DOI: https://doi.org/10.6092/issn.2532-8816/8581

Un'Analisi Multidimensionale della Ricerca Italiana nel Campo delle Digital Humanities e della Linguistica Computazionale

[1]Rachele Sprugnoli, [2]Gabriella Pardelli, [2]Federico Boschetti e [2]Riccardo Del Gratta

[1]Digital Humanities Group, Fondazione Bruno Kessler, Trento – CIRCSE, Università Cattolica del Sacro Cuore, Milano

[2]Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche, Pisa

[1]rachele.sprugnoli@unicatt.it
[2][gabriella.pardelli;federico.boschetti;riccardo.delgratta]@ilc.cnr.it

Abstract. This article proposes the first comparative study of four years of Italian conferences in the fields of Digital Humanities and Computational Linguistics. More specifically, we created a corpus consisting of the contributions presented in the AIUCD and CLiC-it conferences between 2014 and 2017 to which we applied a multidimensional analysis taking into consideration: (i) the study of collaborations between authors using social networks analysis techniques, (ii) the automatic extraction of terminology and information and (iii) the examination of citational practices. By combining both qualitative and quantitative methods of investigation, this paper aims to shed light on convergences and discrepancies between two research areas that historically have common origins.

Questo articolo propone il primo studio comparativo di quattro anni di conferenze italiane nel campo delle Digital Humanities e della Linguistica Computazionale. Nello specifico, è stato creato un corpus costituito dai contributi presentati tra il 2014 ed il 2017 nelle conferenze AIUCD e CLiC-it a cui è stata applicata un'analisi multidimensionale prendendo in considerazione: (i) lo studio delle collaborazioni tra autori usando tecniche di analisi delle reti sociali, (ii) l'estrazione automatica di terminologia ed informazioni e (iii) l'esame delle pratiche citazionali. Combinano metodi di indagine sia qualitativi che quantitativi, questo lavoro vuole far luce su convergenze e discrepanze tra due ambiti di ricerca che storicamente hanno sorgenti comuni.

Introduzione

L'Umanistica Digitale e la Linguistica Computazionale hanno sorgenti comuni[1]. Si preferisce parlare di sorgenti e non di radici, perché le molteplici relazioni che queste discipline hanno intrecciato nel corso del tempo fanno pensare più ad un fiume carsico che ad un albero[2]. Pare non vi sia stata, infatti, una iniziale divisione in aree di competenza e una successiva diramazione in settori specifici. Al contrario, come avviene a un fiume carsico, le linee d'indagine si sono più volte separate e più volte sono tornate a confluire, magari dopo aver seguito percorsi sotterranei basati sull'umile attività di laboratorio verso progetti pilota o sulla semplice comunicazione privata intorno ad idee abbozzate, prima di riacquistare visibilità in grandi progetti finanziati o in riviste e convegni autorevoli.

Se si accetta lo schema proposto da Gibbon,[3] Informatica e Scienze umane stanno alla base della Computational Linguistics, che nel corso degli anni si è settorializzata in Mathematical Computational Linguistics, Natural Language Processing (NLP) e Computational Corpus Linguistics (CL) e, ibridandosi nuovamente con le Scienze umane di tipo tradizionale, in Humanities Computing (HC). Le Digital Humanities[4] (DH), secondo questo schema, nascerebbero dalla confluenza di alcuni settori NLP e CL con HC. Ma, al di là degli atti di nascita di monografie, riviste e convegni internazionali che hanno sancito l'una o l'altra titolatura, è fuor di dubbio che temi di linguistica computazionale e temi di umanistica digitale sono stati trattati dai medesimi autori, attraverso le medesime pubblicazioni e nei medesimi convegni fin dal loro primo apparire 56..

È in questo contesto che nasce l'idea di confrontare gli andamenti delle conferenze annuali dell'Associazione di Informatica Umanistica e Culture Digitali (AIUCD) e dell'Associazione Italiana di Linguistica Computazionale (AILC, che organizza la Conferenza Italiana sulla Linguistica Computazionale, CLiC-it), nelle quattro annate (2014-2017) di coesistenza delle associazioni, per le quali sono disponibili i contributi. Entrambe sono associazioni giovani, nate in un momento storico in cui gli studiosi italiani con competenze ibride fra ambiti umanistici e scientifici sentono il bisogno di rivendicare la propria identità e autonomia disciplinare e di partecipare attivamente alla propria comunità scientifica di riferimento, non solo tramite l'ordinaria attività accademica di docenza e pubblicazione, ma anche attraverso l'organizzazione di eventi aggregativi e la presenza nei gruppi che trovano espressione sui social media 32..

1 Non a caso Padre Busa è riconosciuto come fondatore di entrambe, essendosi occupato, fin dalla fine degli anni Quaranta del secolo scorso, di problemi di linguistica computazionale (come la creazione di indici e concordanze) applicata a testi di interesse umanistico (ovvero gli scritti di San Tommaso d'Aquino).

2 La contrapposizione fra fiume carsico e albero è presa in prestito evidentemente da Maas 62., anche se il filologo in quel passo è più interessato a come l'acqua del fiume (cioè la tradizione) nei suoi mille rivoli si contamini di elementi spuri, che non al modo in cui i diversi rami si ricongiungano e tornino a separarsi e riunirsi ancora.

3 http://wwwhomes.uni-bielefeld.de/gibbon/ELKL-4/LangTecLearningfromEndangeredLang03.mini.pdf

4 Si rimanda a Terras et al. 61. per tracciare il passaggio da Humanities Computing a Digital Humanities, collocabile intorno a pubblicazioni e convegni del 2005.

Pensiamo che questo studio comparativo, che avvia il monitoraggio delle conferenze, possa essere utile per mettere in vista risorse e pratiche comuni alle due associazioni, in modo da rafforzare la sinergia e la condivisione di risorse fra le due comunità, o individuare discrepanze che potrebbe essere interessante colmare.

Dopo una panoramica sullo stato dell'arte inerente l'analisi di comunità scientifiche da vari punti di vista, vengono illustrati i metodi e gli strumenti adottati per confrontare le informazioni delle due serie quadriennali di conferenze. Le analisi condotte sono poi descritte nel dettaglio: tali analisi hanno per oggetto le comunità (autori, genere, affiliazione, partecipazione congiunta ad entrambe le conferenze), i contenuti testuali degli articoli (estrazione terminologica in diacronia) e le voci citazionali (misura di categorie documentali, censimento delle associazioni coinvolte nella pubblicazione). A seguire, nelle conclusioni sono formulate delle ipotesi interpretative che lasciano spazio a future indagini.

Stato dell'arte

Negli studi sulla produzione di contributi pubblicati su riviste o presentati a conferenze scientifiche si possono distinguere principalmente tre linee d'indagine: l'analisi delle comunità, l'analisi dei contenuti e l'analisi citazionale.

L'analisi delle comunità ha come obiettivo la descrizione delle relazioni che intercorrono fra i singoli autori e i loro co-autori, lo studio delle affiliazioni, l'osservazione del rispetto o meno della parità di genere nella distribuzione degli autori, etc. L'analisi delle comunità si basa quindi sui metadati degli articoli, abitualmente presenti nell'intestazione e si avvale sempre più spesso dei metodi adottati tradizionalmente per la social network analysis 23.. L'aspetto più delicato consiste nell'individuazione delle caratteristiche (*features*) poste sotto osservazione, in prospettiva sincronica o in prospettiva diacronica, come già si vede in White e McCain 51., che analizzano dati relativi a riviste di informatica, distribuiti nell'arco di ventiquattro anni. Fra le *features* più rilevanti si trovano la nazionalità dei singoli autori, la composizione internazionale dei gruppi di ricerca cui appartengono i co-autori, la natura (nazionale o internazionale) dei fondi che hanno finanziato la ricerca. Una simile strutturazione dell'analisi emerge in vari altri contributi sull'argomento, fra cui Münster e Ioannides 47. per conferenze dedicate alla Digital Heritage e Mariani et al. 46., Del Gratta et al. 57., Bartolini et al. 58. per LREC, il più importante convegno annuale sulle risorse linguistiche.

L'analisi dei contenuti ha invece come obiettivo l'individuazione degli argomenti (*topics*) più rilevanti trattati nelle pubblicazioni sotto osservazione, nell'estrazione della terminologia di dominio e nello studio delle variazioni in prospettiva diacronica (*trends*). L'analisi dei contenuti si può applicare ai titoli, alle keyword, agli abstract, ma soprattutto al corpo degli articoli. Johri et al. 44. applicano tecniche di topic modelling per studiare, nell'ambito della Linguistica Computazionale, interessi emergenti, stabili e in declino dal 1965 al 2009. Dunaiski et al. 42. applicano tecniche di keyphrase extraction ai titoli e agli abstract di articoli scientifici contenuti

nella ACM Digital Library mentre Mariani et al. 46. estraggono ed analizzano diacronicamente la frequenza di termini rilevanti da 15 anni di atti della conferenza LREC. Infine, mentre Radev et al. 59. sperimentano l'uso di algoritmi di classificazione di testi per strutturare l'antologia della ACL (Association for Computational Linguistics), Wang 60. costruisce una rete di co-occorrenza delle keyword selezionate dagli autori negli articoli di Digital Humanities presenti sulla piattaforma Web of Science.

Relativamente ad analisi sia delle collaborazioni che del contenuto, è importante citare la piattaforma SAFFRON[5] che si pone a cavallo tra le due modalità di analisi. Gli algoritmi implementati in SAFFRON permettono l'estrazione di conoscenza dai testi degli articoli scientifici in modalità completamente automatica. Nello specifico, vengono estratti topics, keywords e autori e viene poi creata una rete di connessioni tra questi elementi per individuare i maggiori esperti per ciascun argomento estratto 36..

Passando all'analisi citazionale, questa ha come obiettivo il monitoraggio dei tipi di materiali citati nella bibliografia degli articoli (monografie, articoli su rivista, atti di convegni, etc.), delle relazioni fra i documenti citati e l'articolo che li cita (autocitazioni, atti di un medesimo convegno, etc.), delle co-citazioni. Wolfe Thompson 50. rileva come, in ambito umanistico, la monografia e la citazione delle fonti primarie rivestano un ruolo centrale rispetto ad altri ambiti. Radev et al. 59., nel presentare l'ACL Anthology Network, indicano quali siano gli autori più citati e come siano organizzate le reti di collaboratori/co-autori. Wang 60., tra i vari parametri bibliometrici, tiene conto anche della distribuzione delle lingue diverse dall'inglese. Tang et al. 48., a tal proposito, osservano che nell'ambito delle Digital Humanities la collaborazione fra co-autori è fortemente condizionata dalla lingua e dai confini geografici. Taskin e al. 49. rilevano poi come la valutazione delle citazioni richieda l'impiego di metodi qualitativi da affiancare ai metodi quantitativi, al fine di osservare la reale rilevanza e gli scopi della citazione, oltre al numero di volte in cui un autore è citato.

Prendendo come punto di partenza i lavori sopra citati, in questo studio abbiamo voluto adottare una prospettiva multi-dimensionale considerando sia l'analisi delle reti di co-autori, che l'estrazione di informazione dal testo dei contributi, che la disamina delle citazioni. Abbiamo quindi combinato varie metodologie di ricerca ed indagine sia qualitative che quantitative per far luce sulle caratteristiche e le dinamiche interne delle comunità di Digital Humanities e Linguistica Computazionale in Italia.

Metodologia

Questo articolo prende in considerazione tre aspetti che aiutano a definire le comunità oggetto del nostro studio nelle loro somiglianze e differenze: la collaborazione tra autori, i contenuti rilevanti o innovativi anche in prospettiva diacronica, le citazioni che stabiliscono connessioni con altri lavori o altre discipline. Questi tre aspetti sono stati indagati identificando le unità di analisi più adatte alla scopo ed adottando metodi specifici:

5 http://saffron.insight-centre.org/

1) lo studio delle collaborazioni ha come unità di analisi gli autori e le loro proprietà principali estrapolate dai metadati dei contributi. Per studiarne le caratteristiche si sono adottati i principi dell'analisi delle reti sociali calcolando metriche standard e creando visualizzazioni importando matrici manualmente compilate in Gephi 24.;
2) i contenuti rilevanti ed innovativi sono stati estratti dal corpus formato dai contributi pubblicati negli atti delle conferenze nei 4 anni considerati. I testi sono stati processati usando strumenti di trattamento automatico del linguaggio allo stato dell'arte, come CoreNLP 1. e Tint 3., e di corpus analysis, come AntConc 20..
3) le citazioni sono state estratte manualmente dalle sezioni bibliografiche dei contributi, accorpate in macrocategorie secondo tipologie documentarie e analizzate in modo statistico/comparativo.

La sezione successiva descrive nel dettaglio sia la metodologia che i risultati per ciascun tipo di analisi. Per motivi di spazio alcune delle visualizzazioni risultanti dalle analisi condotte non sono state riportate nell'articolo ma tutte sono disponibili su un sito online appositamente creato: www.resourcebook.eu/trends/analisiAIUCDCLICit.html.

Analisi

Nelle prossime sotto-sezioni verranno descritti nel dettaglio i metodi ed i risultati ottenuti dai tre tipi di analisi scelti per questo studio.

Analisi delle Comunità

Al fine di descrivere le comunità di AIUCD e CLiC-it, abbiamo applicato l'analisi delle reti sociali ("social network analysis") alle relazioni tra co-autori 21.. In ambito bibliometrico le reti di co-autori sono utilizzate per indagare la struttura e le caratteristiche delle collaborazioni scientifiche 23., 22.: tali reti vengono anche chiamate "grafi di collaborazione" 31.. Basandoci su questo tipo di approccio, abbiamo adottato un modello di rete binaria non orientata da cui abbiamo estratto una serie di dati quantitativi e di metriche. La rete è stata costruita sulla base della co-occorrenza dei nomi degli autori di ogni contributo per tutti gli articoli delle conferenze AIUCD e CLiC-it tra il 2014 ed il 2017. Ogni nodo della rete corrisponde ad un autore mentre ogni arco rappresenta una relazione di co-autorialità. Il numero di contributi a cui gli autori hanno collaborato è rappresentato con archi a differente spessore (i.e. archi pesati), la scala di spessori è proporzionale al numero e più un arco è spesso e più è forte la relazione di co-autorialità. Per ogni nodo sono state aggiunte proprietà relative a genere[6] (donna o uomo), affiliazione, città e nazionalità dell'affiliazione, appartenenza ad un ente senza scopo

6 Siamo consapevoli del fatto che assegnare un genere ad un autore senza il suo esplicito consenso sia una semplificazione che solleva problemi etici 30., tuttavia è un processo necessario per fornire dati sulla sotto-rappresentazione delle donne nelle comunità oggetto di questo studio.

di lucro (ad esempio, università, centri di ricerca privati, scuole, biblioteche, ONG) o ad un'azienda.[7] I dati, manualmente compilati, sono stati importati in Gephi[8] per la visualizzazione delle reti ed il calcolo delle metriche.

Tabella 1 riassume le caratteristiche principali delle reti nelle varie edizioni di AIUCD e CLiC-it in una prospettiva diacronica. Dopo la prima edizione di AIUCD il numero di autori è più che raddoppiato ma comunque minore del numero di autori di CLiC-it (con la sola eccezione di CLiC-it 2015 che ha registrato il numero minimo di autori, i.e. 129). La percentuale di donne tra gli autori è un dato importante: la rappresentanza di genere nelle conferenze scientifiche è un problema molto sentito 25. ed i concetti di diversità ed inclusione sono alla base di studi e dibattiti sia nel campo delle DH che della linguistica computazionale[9] (26., 28.). In CLiC-it la percentuale di donne autrici di contributi ha avuto una leggera crescita nelle prime 3 edizioni ma un calo del 2,88% nel 2017: la percentuale di 37,66% è comunque superiore rispetto al 33% registrato negli articoli dell'intera antologia dell'associazione di linguistica computazionale (ACL) 27.. Anche in AIUCD si registra una percentuale di donne superiore a quella di altre conferenze del campo: ad esempio, Weingart 29. riporta una percentuale del 34,6% nella conferenza internazionale DH 2015. Tuttavia, benché le prime due edizioni abbiano avuto un perfetto bilanciamento tra i generi, questo equilibrio non si è ripetuto nelle edizioni più recenti. Nel corso degli anni sia AIUCD che CLiC-it hanno attirato un sempre maggior numero di stranieri: particolarmente rilevante è la netta crescita di autori non italiani in AIUCD 2017 (45.25%), favorita dall'organizzazione congiunta di altri due eventi di respiro internazionale, l'EADH day e il workshop Dixit. La tendenza relativa alla partecipazione di aziende risulta invece avere un andamento opposto nelle due conferenze: è in forte calo in AIUCD (con una diminuzione dell'11,95% tra la prima e l'ultima edizione considerata) ed in crescita in CLiC-it (con un aumento dell'8,2% tra la prima e l'ultima edizione).

	AIUCD	AIUCD	AIUCD	AIUCD	CLiC-it	CLiC-it	CLiC-it	CLiC-it
	2014	2015	2016	2017	2014	2015	2016	2017
nodi	49	132	101	137	167	129	148	154
archi	79	187	137	268	249	214	236	254
donne	51,02%	50%	35,64%	40,15%	39,52%	39,53%	40,54%	37,66%

7 È importante notare che i cambi di affiliazione occorrono in maniera marginale nei dati presi in esame: solo il 2,8% degli autori ha cambiato affiliazione tra il 2014 ed il 2017. Inoltre solo lo 0.9% si è trasferito dall'Italia all'estero o viceversa e solo lo 0.4% si è trasferito da un ente senza scopo di lucro ad un'azienda, o viceversa.

8 https://gephi.org/

9 Si veda anche il Workshop on Women and Underrepresented Minorities in NLP (WiNLP) poi ribattezzato Widening NLP: http://www.winlp.org/

stranieri	14,28%	15,91%	33,66%	45,25%	14,37%	17,05%	15,54%	21,43%
aziende	16,33%	10.,61%	4,95%	4,38%	4,79%	0,78%	4,05%	12,99%

Tabella 1: Dati quantitativi relativi alle reti delle varie edizioni di AIUCD e CLiC-it.

Le reti relative a ciascuna edizione sono state poi accorpate in un'unica matrice per rappresentare le comunità di autori di AIUCD e CLiC-it nel loro complesso.[10] Tabella 2 riporta dati e metriche di queste reti globali. Il maggior numero di nodi di CLiC-it rispetto ad AIUCD indica un maggior numero di autori coinvolti nella comunità. Per quanto riguarda gli archi, benché il loro numero indichi la quantità di collaborazioni tra autori, da solo non è sufficiente per misurare il livello e l'intensità di tali collaborazioni; a questo scopo abbiamo calcolato delle metriche specifiche. La *densità* è un indicatore del livello di connettività della rete: è calcolata come il rapporto tra il numero di archi effettivi tra nodi ed il numero di tutti gli archi possibili (ovvero il caso in cui ogni autore è collegato a ogni altro autore della rete). La densità è uguale per entrambe le reti e molto bassa (0,01) indicando una rete non densa, con un'alta selettività nella scelta dei propri co-autori. Il *grado medio* è un'altra metrica standard che misura la coesione e corrisponde alla media del numero degli archi per ogni nodo; in altre parole, indica il numero medio di co-autori per autore. Il valore di AIUCD (3,56) è poco minore di quello misurato sugli articoli in SCOPUS nel campo DH 48. mentre il valore di CLiC-it (3.91) è maggiore di quello delle principali conferenze di linguistica computazionale (3.79 in ACL, 3.21 in Coling, 3.79 in EMNLP) ma molto minore che in LREC (6,42) 46.. Il numero di *componenti connessi* corrisponde al numero delle parti della rete che non sono connesse tra di loro: un valore alto come in AIUCD (122) indica che la rete è molto più frammentata, con poco scambio tra gli autori, rispetto a quella di CLiC-it.

Le strutture delle reti di co-autori in ambito scientifico tendono ad avere un'organizzazione detta di *piccolo-mondo* ("small-world"): questo tipo di struttura caratterizza la collaborazione in molte comunità, ad esempio in fisica, in biologia, in matematica e in sociologia 33. 34.. Il recente lavoro di Tang et al. 48. ha invece definito la struttura della comunità internazionale DH come una rete "plural-world" 35. a causa della sua frammentazione. L'alto numero dei *componenti connessi* sembrerebbe assegnare quest'ultimo tipo di struttura anche ad AIUCD ma per confermarlo è necessario esaminare le caratteristiche della componente principale del grafo, ovvero della sotto-rete con il maggior numero di nodi interconnessi, riportate in Tabella 3. Il basso numero di nodi della componente principale in AIUCD, corrispondente a meno del 20% del totale dei nodi della rete, unito all'alto coefficiente di clustering medio (0.86), che ha un valore molto più elevato rispetto a quello che sarebbe il coefficiente di una rete casuale creata con lo stesso numero di nodi (0.03), conferma, anche per l'Italia, la struttura di tipo "plural-world", fatta di frammenti sub-disciplinari distinti benché strettamente organizzati su argomenti, problemi o metodi comuni. Questa struttura non è però adatta a descrivere CLiC-it che, al contrario, ha un'estesa componente principale corrispondente al 41,32% dei nodi totali

10 Le reti globali di AIUCD e CLiC-it sono disponibili, in formato interattivo, su www.resourcebook.eu/trends/analisiAIUCDCLICit.html

della rete.

	AIUCD	CLiC-it
nodi	361	409
archi	642	799
densità	0,01	0,01
grado medio	3,56	3,91
componenti connessi	122	64

Tabella 2: Dati e metriche caratterizzanti le matrici globali di tutte le edizioni di AIUCD e CLiC-it.

	AIUCD	CLiC-it
componente principale (# nodi)	70	169
componente principale (% nodi)	19,39%	41,32%
coefficiente di clustering medio	0,86	0,81
(valore random previsto)	0,03	0,05
lunghezza media del percorso	3,68	4,72
(valore random previsto)	2,04	2,62

Tabella 3: Metriche relative alla componente principale estratto dalle reti globali di AIUCD e CLiC-it.

Le componenti principali di AIUCD e CLiC-it sono visualizzate in Illustrazione 1 e Illustrazione 2.[11] Sono colorati in rosa gli autori di nazionalità italiana ed in verde quelli di nazionalità straniera per evidenziare le collaborazioni internazionali. In AIUCD solo 7 autori (corrispondenti al 10% dei nodi della componente principale) sono affiliati ad istituzioni fuori confine, nello specifico in Francia, Croazia, Canada, Germania e Gran Bretagna. Nella componente principale di CLiC-it gli autori con affiliazione non italiana sono 25 (14.8%) provenienti in particolare da Francia e Paesi Bassi ma anche, in percentuale minore, da Cina, Germania, Gran Bretagna, Svezia, Stati Uniti e Spagna.

11 Per questa analisi abbiamo considerato l'ultima affiliazione registrata da ciascun autore.

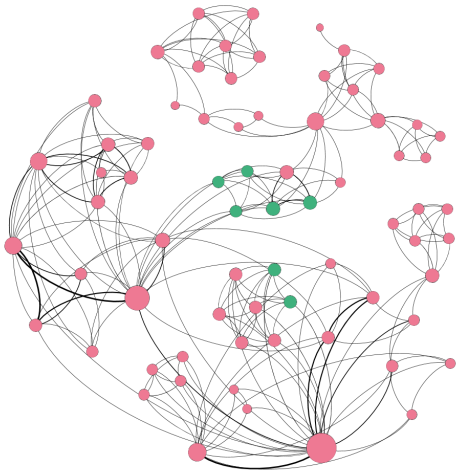

Illustrazione 1: Componenti principali di AIUCD nella rete di co-autori. I nodi rappresentano gli autori e sono colorati in base alla nazionalità italiana o meno delle loro affiliazioni (rosa se italiana, verde se straniera).

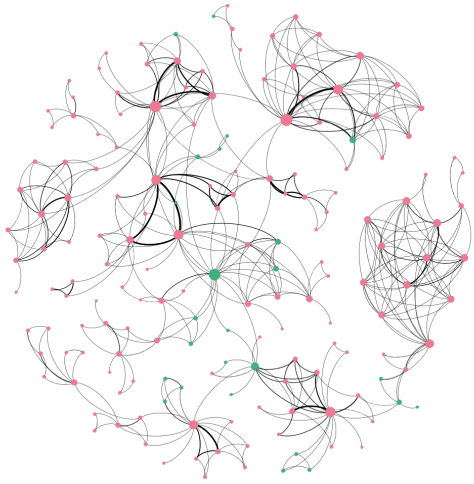

Illustrazione 2: Componenti principali di CLiC-it nella rete di co-autori. I nodi rappresentano gli autori e sono colorati in base alla nazionalità italiana o meno delle loro affiliazioni (rosa se italiana, verde se straniera).

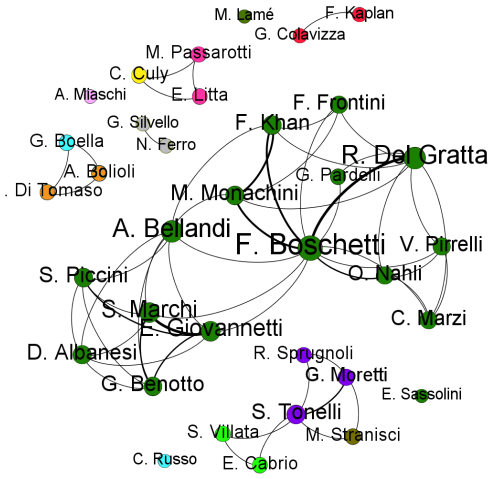

Illustrazione 3: Rete dei 35 autori che hanno pubblicato sia nella conferenza AIUCD che in CLiC-it e delle loro inter-relazioni.

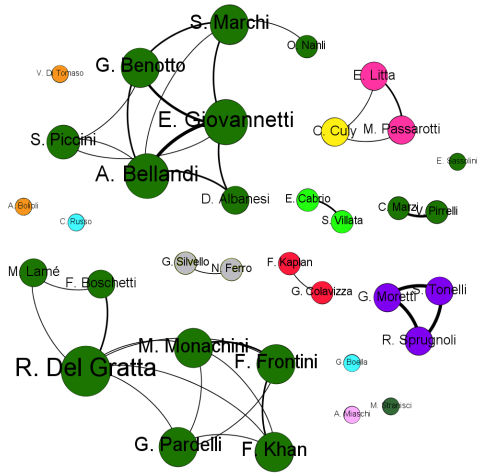

Illustrazione 4: Rete dei 35 autori che hanno pubblicato sia nella conferenza AIUCD che in CLiC-it e delle loro inter-relazioni.

Un'analisi incrociata degli autori che hanno pubblicato in entrambe le conferenze ha evidenziato la presenza di 35 nodi comuni tra le due reti. Le connessioni tra questi autori sono visualizzate in Illustrazione 3 e Illustrazione 4 dove il colore dei nodi corrisponde alla loro affiliazione. I 35 autori sono affiliati a 13 diverse istituzione. Le istituzioni principali sono: l'Istituto di Linguistica Computazionale del CNR a cui appartiene quasi la metà degli autori (45,71%, in verde), seguito dalla Fondazione Bruno Kessler (8.57%, in viola), l'Università di Padova, CELI, EPFL, Università Cattolica e Università di Torino (tutte al 5,71% e rispettivamente in azzurro, arancione, rosso, fucsia e grigio chiaro). Le altre istituzioni hanno una percentuale di occorrenza minore del 3%. Questo gruppo di autori comuni ad AIUCD e CLiC-it comprende un'azienda, CELI, e alcune istituzioni straniere come l'EPFL di Losanna. È interessante notare che la maggior parte delle collaborazioni tra questi autori si instaura all'interno della stessa istituzione, soprattutto in CLiC-it. In AIUCD ci sono invece collaborazioni tra istituzioni diverse, per esempio quella tra Boella, Bolioli e Di Tomaso ha una connotazione geografica precisa visto che le loro istituzioni hanno sede a Torino, ma anche tra nazioni diverse come tra Italia e Stati Uniti (Passarotti, Litta e Culy) e tra Italia e Francia (Tonelli, Cabrio e Villata).

Analisi del Contenuto

L'analisi del contenuto si è basata sullo studio del corpus di contributi presentati alle conferenze AIUCD e CLiC-it negli anni di interesse (dal 2014 al 2017). I dati quantitativi relativi al corpus raccolto ed analizzato sono riportati in Tabella 4: il numero dei token comprende titolo e testo del contributo includendo abstract, note e didascalie di figure e tabelle ma escludendo le sezioni bibliografiche. Il corpus è stato diviso in due sotto-corpora (AIUCD e CLiC-it).

Una delle caratteristiche principali del corpus è quella di contenere sia contributi in italiano che in inglese: le due lingue sono considerate lingue ufficiali in entrambe le conferenze per aumentare l'inclusività e la partecipazione internazionale. I testi inglesi sono stati tokenizzati e lemmatizzati con Stanford CoreNLP[12] mentre per i testi italiani abbiamo usato Tint.[13] Il forte aumento dei contributi in inglese in AIUCD 2017 è in linea con l'alta percentuale di autori stranieri registrati quell'anno (45.25%, si veda Tabella 1). In CLiC-it, il numero di contributi in italiano è in costante diminuzione giungendo a coprire solo il 7,4% dei lavori presentati nel 2017.

Per quanto riguarda il numero di token, è molto evidente la differenza tra le due conferenze: tale diversità è dovuta al fatto che in AIUCD, ad eccezione del 2014, sono stati pubblicati degli abstract fino a 500 parole nel 2015 e fino a 1000 negli anni successivi. CLiC-it richiede invece la presentazione di un articolo di 4 pagine corrispondenti ad una media di 2000 parole a contributo.

12 https://stanfordnlp.github.io/CoreNLP/

13 http://tint.fbk.eu/

		AIUCD	AIUCD	AIUCD	AIUCD	CLiC-it	CLiC-it	CLiC-it	CLiC-it	TOT
		2014	2015	2016	2017	2014	2015	2016	2017	
FILE	EN	17	12	13	46	62	46	47	54	
	IT	1	41	32	16	13	6	8	4	
TOT		18	53	45	62	75	52	55	58	418
TOKEN	EN	68.756	5.848	16.859	57.099	152.290	96.631	118.314	148.058	
	IT	10.403	18.210	36.774	29.457	31.082	12.321	20.216	11.524	
TOT		79.159	24.058	53.633	86.556	183.372	108.952	138.53	159.582	709.165

Tabella 4: Composizione del corpus: numero di file e di token divisi per conferenza e per anno. Nell'analisi non sono stati inclusi un contributo di AIUCD 2015, in quanto si configurava come brochure promozionale di un'azienda e non come abstract scientifico, ed un contributo in tedesco di AIUCD 2016.

La prima analisi effettuata ha riguardato l'identificazione di temi e metodi innovativi per ciascuna edizione considerata attraverso l'estrazione dei nuovi termini introdotti di anno in anno negli atti delle conferenze. A questo scopo, i lemmi dei contributi di ogni anno sono stati confrontati con quelli delle edizioni precedenti ed il risultato è stato ispezionato manualmente per controllare l'eventuale presenza degli stessi termini sia in inglese che in italiano (e.g. *collation*/*collazione*). L'ispezione manuale ha anche permesso di selezionare un set di termini particolarmente interessanti presentati in Tabella 5. Al di là di *linking* nel 2015 e *collation* nel 2016, introdotti nello stesso anno in entrambe le conferenze, i nuovi termini elencati nella Tabella rivelano un interesse per temi e tecnologie differenti o una diversa tempistica della loro introduzione nella comunità. In CLiC-it è particolarmente evidente la sempre maggior attenzione alle reti neurali, per cui vengono nominate reti di diverso tipo come *LSTM* (Long Short-Term Memory), *bi-LSTM* (bidirectional LSTM) e *RNN* (Recurrent Neural Network) ma anche software specifici (*Tensorflow*, *Theano*, *Keras*) e applicazioni come la neural machine translation (*NMT*). Il word embedding e la semantica distribuzionale, due temi presenti fin dal 2014 in CLiC-it, vengono introdotti in AIUCD nel 2016 quando appaiono per la prima volta i termini *distributional*, *distribuzionale*, *embeddings*. Negli atti della stessa edizione compare anche l'espressione *scuola-lavoro*: l'alternanza scuola-lavoro era stata resa obbligatoria a partire

dall'anno scolastico 2015/16 e le iniziative ad essa legate erano diventate fin da subito oggetto di contributi in AIUCD ma non in CLiC-it. È poi interessante notare che le grandi infrastrutture europee legate al mondo DH, i.e. *Dariah* e *Clarin*, vengono menzionate per la prima volta a CLiC-it solo nel 2015 benché l'area tematica denominata "Natural Language Processing for the Digital Humanities" fosse stata istituita già per la prima edizione della conferenza nel 2014.

2015	AIUCD	user-oriented, user-orientation, fabrication, immersive/immersivo, linking
2015	CLiC-it	argumentation, emoji, linking, profiling, Dariah, Clarin
2016	AIUCD	collation/collazione, distributional/distribuzionale, embeddings, scuola-lavoro
2016	CLiC-it	collation, gamification, LSTM, bi-LSTM, RNN, Tensorflow, Theano, Keras
2017	AIUCD	partecipative, polarity/polarità, bigram, skip-gram, n-gram
2017	CLiC-it	NMT, word2phrase, glove, code-mixing, neo-latin, adjudication, reconciliation

Tabella 5: Esempi di nuovi termini introdotti di anno in anno nelle conferenze.

La seconda fase dell'analisi contenutistica ha riguardato l'estrazione degli argomenti principali caratterizzanti i due sotto-corpora. Diversi approcci bibliometrici si basano sullo studio delle parole-chiave scelte dagli autori stessi al momento dell'invio del contributo, si vedano tra gli altri 3. e 4., tuttavia tale informazione non era in nostro possesso. In altri lavori, invece, viene applicato il topic modeling (6., 5.) per estrarre automaticamente gruppi di parole co-occorrenti in grandi collezioni di articoli (8., 7.). Il topic modeling ha però alcuni punti deboli tra cui il fatto di non integrare informazione linguistica, essendo completamente statistico, e di non permettere l'estrazione di espressioni multi-token. Al fine di superare queste limitazioni ed ottenere delle parole-chiave più linguisticamente motivate e complesse, abbiamo deciso di utilizzare KD, *Key-phrase Digger* 9.. KD identifica concetti-chiave formati da una o più parole in base ad una lista di pattern di parti del discorso specifica per ogni lingua trattata (ad esempio, Aggettivo-Nome per l'inglese e Nome-Aggettivo per l'italiano) e combina questa informazione linguistica con informazione statistica (ad esempio, frequenza e term frequency–inverse document frequency) per creare una lista pesata di termini ed espressioni. Inoltre, KD contiene una funzione che riconosce ed assegna un peso particolare agli

Illustrazione 5: I primi 100 concetti-chiave estratti dagli abstract AIUCD. La dimensione del font è proporzionale al peso normalizzato del concetto-chiave.

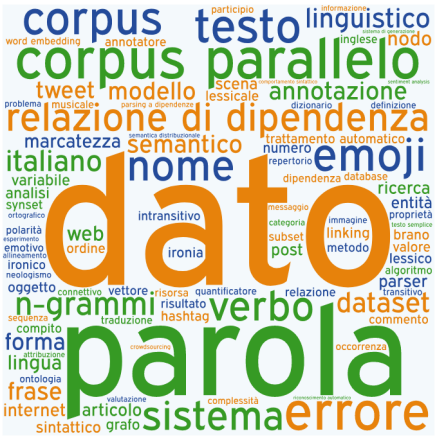

Illustrazione 6: I primi 100 concetti-chiave estratti dai paper CLiC-it. La dimensione del font è proporzionale al peso normalizzato del concetto-chiave.

acronimi non espansi (ad esempio, *XML*) che svolgono un ruolo centrale negli articoli scientifici. Le visualizzazioni riportate in Illustrazione 5 e Illustrazione 6 mostrano il risultato ottenuto con KD su i due sotto-corpora lemmatizzati: i concetti-chiave in inglese sono stati tradotti in italiano, con l'eccezione di espressioni come *word embedding* fortemente codificate in inglese nel linguaggio scientifico, ed il peso è stato normalizzato tra 0 e 1 per rendere il risultato comparabile e non sensibile alla differente lunghezza dei due sotto-corpora. L'estrazione di concetti-chiave multi-token ha reso possibile identificare espressioni molto specifiche e caratterizzanti come *biblioteca digitale*, *patrimonio culturale*, *annotazione semantica*, *authorship attribution*, *fonte primaria* in AIUCD e *corpus parallelo*, *relazione di dipendenza*, *trattamento automatico*, *word embedding*, *sentiment analysis* in CLiC-it. È interessante notare che alcuni dei concetti-chiave di AIUCD in Illustrazione 5 sono specifici di una certa edizione: la loro alta rilevanza globale è legata al tema e al titolo scelto in uno specifico anno della conferenza. Una delle differenze tra la conferenza AIUCD e CLiC-it è infatti che la prima ogni anno ha un titolo che ne identifica il tema principale[14] andando così ad influenzare gli argomenti trattati nei contributi. La presenza, ad esempio, di *museo* e *patrimonio culturale* nella Figura è dovuta al loro alto peso in AIUCD 2015, il cui titolo era "*Digital Humanities e beni culturali: quale relazione?"*, mentre *edizione digitale* e *trascrizione* hanno un peso particolarmente alto in AIUCD 2016, il cui titolo era "*Edizioni digitali: rappresentazione, interoperabilità, analisi del testo e infrastrutture*".

Tra i primi 100 concetti-chiave, 25 sono in comune tra i due sotto-corpora: i.e., *algoritmo, allineamento, analisi, corpus, database, dato, entità, immagine, informazione, internet, lessicale, lessico, lingua, modello, ontologia, oggetto, parola, relazione, ricerca, risorsa, risultato, semantico, sistema, testo, web*. Questi termini, tutti formati da una sola parola e quindi non particolarmente specifici, riguardano, tra gli altri, due delle principali categorie della mente umana 10. su cui si basa l'information science (i.e., *dato* e *informazione*), due dei livelli che compongono il sistema lingua (i.e., *lessico*/*lessicale*, *semantico*), gli oggetti principali dell'analisi linguistica (come *lingua*, *testo*, *parola, corpus*). Su questo ultimo aspetto si noti come AIUCD presti maggiore attenzione all'asse sintagmatico, quindi al *testo*, mentre CLiC-it si focalizzi più sull'asse paradigmatico, quindi sulla *parola*. Interessante notare anche l'attenzione posta su

14 In AIUCD 2014 il titolo era: "La metodologia della ricerca umanistica nell'ecosistema digitale"; in AIUCD 2015 " Digital Humanities e beni culturali: quale relazione?"; in AIUCD 2016 "Edizioni digitali: rappresentazione, interoperabilità, analisi del testo e infrastrutture"; in AIUCD 2017: "Il telescopio inverso: big data e distant reading nelle discipline umanistiche".

internet, come principale infrastruttura di scambio dati, e sul *web* come fondamentale medium comunicativo e risorsa informativa.

Benché i sopra citati termini siano presenti come concetti-chiave sia in AIUCD che in CLiC-it, il loro andamento nel tempo non è stato costante né uguale nelle due conferenze. Illustrazione 8 e Illustrazione 7 mostrano come è variata la rilevanza di alcuni concetti-chiave comuni nel corso delle varie edizioni delle conferenze prendendo in considerazione il loro peso normalizzato. Si noti che, anche in questa analisi, i concetti-chiave in inglese sono stati tradotti in italiano. Il concetto-chiave *corpus* presenta una crescita nelle ultime due edizioni di AIUCD mentre in CLiC-it, negli stessi anni, si ha un calo della sua rilevanza: analizzando altri concetti-chiave simili in CLiC-it si nota una preferenza, negli stessi contesti d'uso, per il termine *dataset* (per esempio, *training corpus/training dataset*; *annotated corpus/annotated dataset*; *corpus collection/dataset collection*) specialmente nel 2017. *Ontologia* ha un picco solo nell'edizione 2015 di CLiC-it mentre è uno dei principali concetti-chiave in tutte le edizioni di AIUCD anche se con un calo nel 2017. *Semantico* ha una tendenza opposta nelle due conferenze: benché in entrambe abbia un trend altalenante, ha un netto aumento di rilevanza in CLiC-it ed un abbassamento in AIUCD dove nel 2017 non sono presenti espressioni frequenti nelle edizioni precedenti, ad esempio *semantic annotation*, *semantic enrichment*, *semantic link*. Infine, *modello* e *sistema* presentano un andamento analogo registrando un aumento di peso, e quindi di rilevanza, in entrambe le conferenze nel corso delle ultime edizioni.

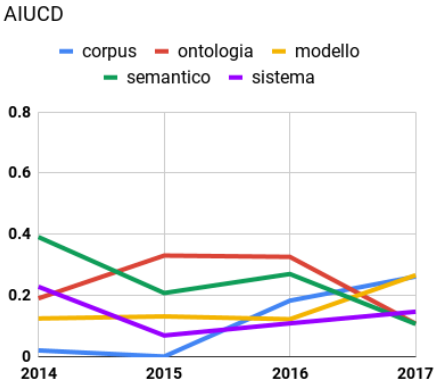

Illustrazione 8: Trend di alcuni concetti-chiave comuni in AIUCD (sinistra) in base al loro peso normalizzato tra 0 e 1.

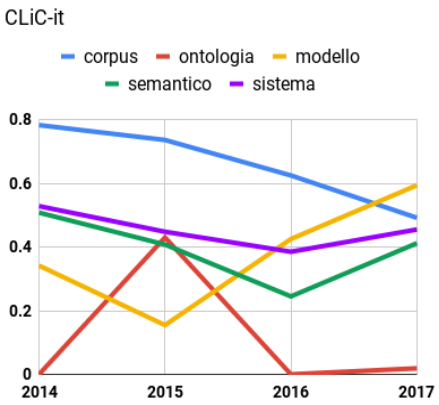

Illustrazione 7: Trend di alcuni concetti-chiave comuni in CLiC-it in base al loro peso normalizzato tra 0 e 1.

Nonostante i concetti-chiave *modello* e *sistema* abbiano tendenze simili nei due sotto-corpora, il loro significato non è perfettamente equivalente in AIUCD e in CLiC-it. In Tabella 6 sono elencati i bigrammi più frequenti per le due conferenze, generati con AntConc,[15] e questo permette di confrontare i principali contesti d'uso dei due concetti-chiave. In CLiC-it sia *modello* che *sistema* fanno riferimento all'idea cardine della linguistica computazionale della lingua come sistema

15 http://www.laurenceanthony.net/software/antconc/

probabilistico descrivibile attraverso la costruzione di modelli computazionali traducibili in programmi eseguibili dal computer, addestrabili e valutabili 11.. Gli stessi concetti-chiave presentano invece una maggiore diversità d'uso in AIUCD. Si fa ad esempio riferimento a sistemi informativi, gestionali, grafici e di comunicazione da un lato e a modelli formali e concettuali dall'altro: la modellazione e la sua formalizzazione sono, non a caso, due aspetti estremamente rilevanti nella ricerca in campo DH (12., 13.).

AIUCD	CLiC-it	AIUCD	CLiC-it
modello/model	modello/model	sistema/system	sistema/system
data model (14) conceptual model (6) modello 3d (5) reference model (5) topic model (5) formal model (4) mathematical model (4) new model (4) theoretical model (4) modello lessicale (3)	language model (66) semantic model (43) topic model (19) computational model (17) nmt model (15) space model (13) regression model (12) trained model (12) bow model (11) distributional model (11)	information system (13) sistema operativo (5) entire system (4) illustrative system (4) management system (4) sistema complesso (3) sistema grafico (3) communication system (3) complex system (3) computational system (3)	ape system (29) mt system (22) best system (19) system performance (16) translation system (15) automatic system (20) sistema asr (12) smt system (12) asr system (11) baseline system (11)

Tabella 6: Primi 10 bigrammi, in termini di frequenza, associati ai concetti-chiave modello e sistema in AIUCD e CLiC-it.

L'analisi del contenuto si completa con uno studio contrastivo tra il sotto-corpus AIUCD e il sotto-corpus CLiC-it usando la funzione `oppose()` di Stylo 14..[16] Tale funzione implementa l'estensione della Zeta di Burrows 16. sviluppata da Craig e Kinney 15. per generare due liste di parole: una contiene le parole significativamente preferite e l'altra quelle significativamente evitate nei contributi di AIUCD rispetto ai contributi di CLiC-it. Illustrazione 9 presenta una visualizzazione delle prime 70 parole delle due liste ordinate in base al valore della Zeta e quindi dalle più alle meno distintive (dall'alto in basso). AIUCD si contraddistingue specialmente per

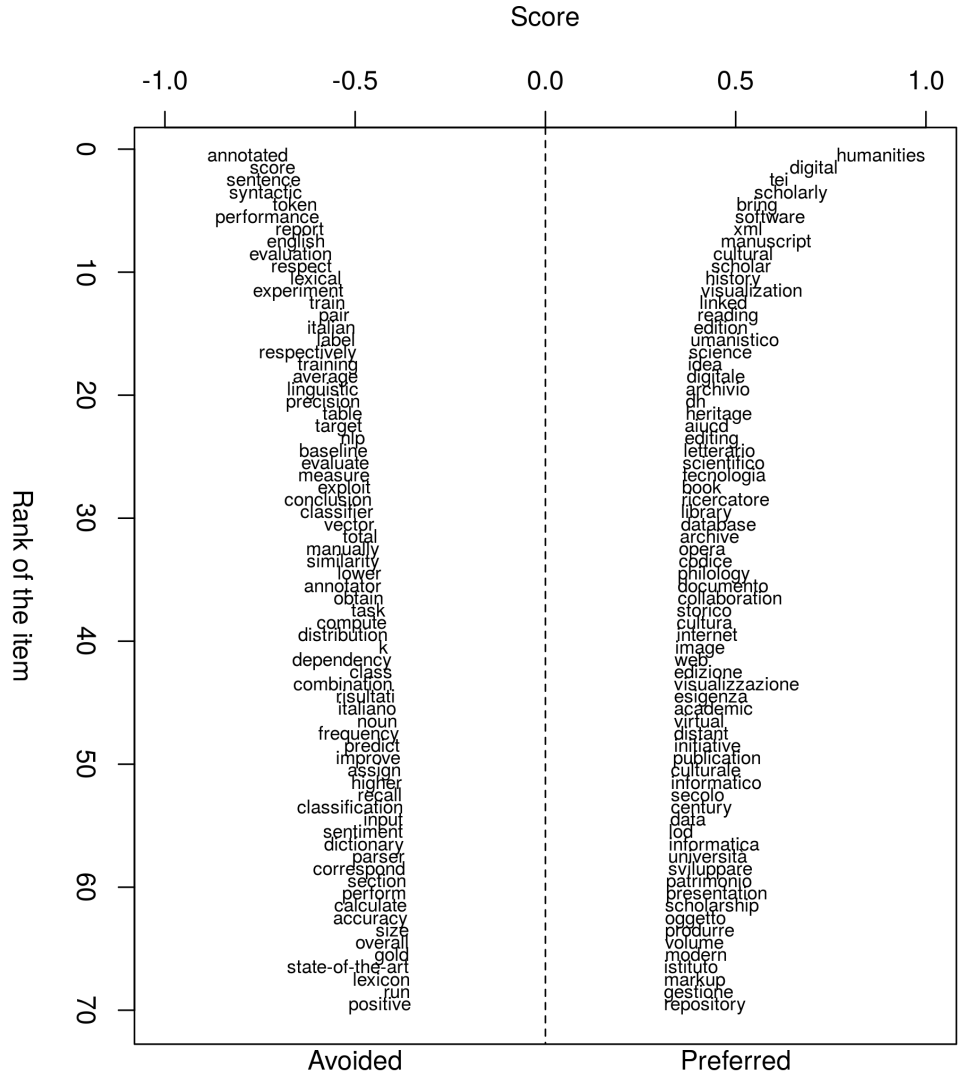

Illustrazione 9: Output della funzione oppose() di stylo, in base alla zeta di Craig, applicata al sotto-corpus AIUCD rispetto al sotto-corpus CLiC-it.

16 https://sites.google.com/site/computationalstylistics/stylo

l'attenzione verso le varie forme di diffusione del sapere (*manuscript*, *edition*/*edizione*, *book*, *volume*) e alle istituzioni ed infrastrutture deputate alla loro raccolta e studio (*archivio*/*archive*, *library*, *università*, *repository*). Al contrario, l'ambito della valutazione, fortemente presente in CLiC-it, è assente in AIUCD benché la necessità di introdurre una pratica di valutazione condivisa in ambito DH sia sempre più sentita (17., 18., 19.): tra le parole evitate troviamo infatti *score*, *performance*, *evaluation*, *precision*, *recall*, *accuracy*.

Analisi Citazionale

Le informazioni necessarie per l'analisi citazionale sono state rilevate dalle voci bibliografiche degli articoli pubblicati negli atti o nei "Book of Abstracts" delle conferenze AIUCD e CLiC-it nell'arco temporale 2014-2017: questa raccolta informativa è andata a costituire il corpus citazionale, comprendente il materiale necessario per l'elaborazione.

Una delle criticità del lavoro è stata rilevata nella mancanza di attenzione ai criteri redazionali e al modello citazionale durante la compilazione delle bibliografie. Tale incuria è giustificata solo in parte dalla complessità delle tipologie documentarie citate nelle due serie di eventi italiani. La difficoltà interpretativa per la corretta annotazione, in alcuni casi, è stata risolta con il ricorso alla rete telematica.

In tutto sono stati esaminati 6.360 riferimenti bibliografici: 1.724 per AIUCD e 4.636 per CliC-it corrispondenti rispettivamente al 27% e al 73% del totale. La maggior grandezza percentuale di CLiC-it rispetto a quella di AIUCD dipende da due elementi: a) maggior numero di articoli per ciascuna conferenza annuale in CLiC-it rispetto a AIUCD; b) comparazione tra due differenti tipologie documentarie, articoli in CLiC-it e abstract in AIUCD.

Il primo studio effettuato sulle voci bibliografiche dei contributi ha riguardato l'esame dettagliato delle associazioni di riferimento. Tale esame ha messo in evidenza l'importante ruolo di editore svolto dalle Associazioni, sia a livello nazionale che internazionale per i diversi ambiti scientifici. Per quanto concerne le associazioni condivise (si veda Illustrazione 10 per informazioni sull'ordine di grandezza dei riferimenti),[17] l'ambito informatico emerge abbondantemente in entrambe le conferenze ed è condiviso con l'Association for Computer Machinery ACM, l'IEEE Computer Society, l'Association for the Advancement of Artificial Intelligence AAAI; l'ambito linguistico e linguistico-computazionale è invece accomunato all'Association for Computational Linguistics ACL, all'European Language Resources Association ELRA, all'Association pour le Traitement Automatique des Langues ATALA, all'European Association for Lexicography EURALEX, alla Linguistic Society of America LSA, alla Global WordNet Association GWA. Le associazioni non condivise tra AIUCD e CLiC-it marcano da vicino le due aree come ad esempio l'Associazione degli Italianisti AdI per AIUCD e l'Associazione Italiana di Intelligenza Artificiale AI*IA per CLiC-it. Questa varietà delle

17 Il nome esteso delle associazioni presenti in Illustrazione 10 è disponibile online insieme ad una serie di grafici aggiuntivi: www.resourcebook.eu/trends/analisiAIUCDCLICit.html

associazioni scientifiche recuperate dalle citazioni, anche soltanto come hapax, avvalora e conferma la multidisciplinarità in entrambe le aree indagate. Alcune di queste associazioni si dimostrano particolarmente sensibili alle tematiche open access e sono impegnate nello sviluppo di migliori strategie di divulgazione della letteratura scientifica, come ad esempio le stesse AILC e AIUCD in Italia e ACL ed ELRA attraverso l' ACL Anthology: A Digital Archive of Research Papers in Computational Linguistics.

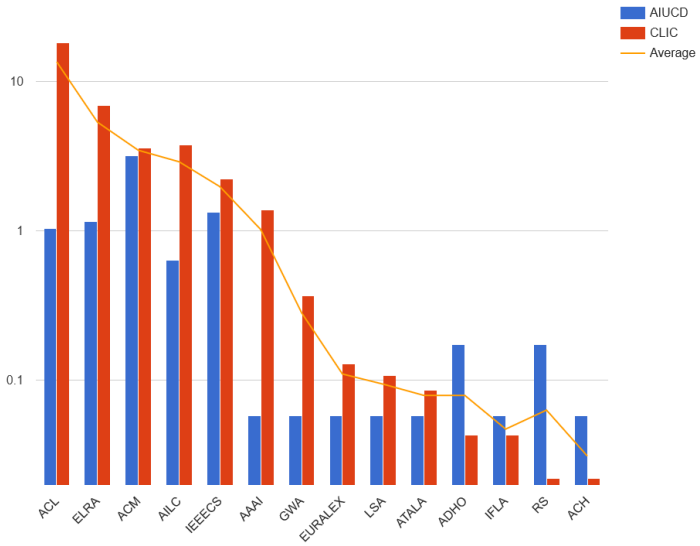

Illustrazione 10: Associazioni comuni tra AIUCD e CLiC-it.

Un altro aspetto della nostra indagine ha riguardato la tipologia di risorse e strumenti proposte nelle bibliografie degli articoli degli autori. I concetti di *risorsa*, *strumento* e *condivisione* sono infatti fondamentali sia per AIUCD che per AILC. Riportiamo a questo proposito le parole raccolte dai siti web di AILC e AIUCD:

"AIUCD diffonde la riflessione metodologica e teorica, la collaborazione scientifica e lo sviluppo di pratiche, risorse e strumenti condivisi nel campo dell'informatica umanistica e nell'uso delle applicazioni digitali in tutte le aree delle scienze umane.".[18]

"AILC promuove e diffonde la riflessione metodologica, teorica e sperimentale, la collaborazione scientifica e lo sviluppo di pratiche, risorse e strumenti condivisi.".[19]

Nei due estratti sopra riportati incontriamo la stessa espressione: *risorse* e *strumenti condivisi*.

Il panorama citazionale del corpus AIUCD e CLiC-it è costituito dall'intreccio fra tipologie di fonti cartacee ed elettroniche. Le fonti cartacee sono rappresentate in larga misura da citazioni a libri e riviste; le edizioni recenti affiancano al mezzo cartaceo anche quello elettronico riconoscibile sia dal codice DOI, Digital Object Identifier, sia dal link inserito dall'autore della citazione per il recupero del testo. Le fonti elettroniche delle citazioni si possono distribuire su

18 http://www.aiucd.it/
19 http://www.ai-lc.it/it/?v=cd32106bcb6d

due ordini:

1) Linkografie a infrastrutture come biblioteche; formati di file creati con Adobe; homepage di siti web di istituzioni scientifiche nazionali e internazionali; pagine web per il download di software/tool; risorse nel settore delle tecnologie della lingua come corpora, dizionari, ontologie, parser, treebank, linee guida/standard, create all'interno di progetti nazionali o europei; pre-prints su archivi aperti per la comunicazione scientifica come ad esempio *arXiv.org*[20] e *CoRR, Computing Research Repository.*[21]

2) Linkografie a social media come blog, posts e tweets, etc. A tal proposito riportiamo le parole di Marcus Banks 37.: "the most profound change lies in the ability for anyone to post "user-generated content" such as blog posts or YouTube clips. Web 2.0 tools are also beginning to influence scientific debate. Blogs are now an established part of the information landscape; they are scrolling public diaries that usually allow comments.".

Queste nuove forme di voci citazionali, pur non essendo forme tradizionali di pubblicazione accademica, sono flussi informativi che contribuiscono ai processi di condivisione della conoscenza. Nelle citazioni recuperiamo anche forme di comunicazione audiovisive come video e televisione. Nella lista che segue sono elencati alcuni esempi di voci a social media e programmi televisivi estratti da AIUCD:

- Video: http://www.egs.edu/faculty/giorgio-agamben/videos/what-is-a-dispositive
- Social media: http://twitter.com/flowchainsensei/status/408167162344648704
- Programma TV: http://www.bbc.co.uk/news/technology-28895098
- Home page: http://www.aspenideas.org/session/count-or-die-why-humanities-neednumbers-survive

L'approccio di tipo quantitativo per l'analisi delle citazioni ha previsto due fasi:

1. inclusione di ciascuna voce in una tipologia documentaria censita in lingua inglese; abstract, book, book chapter, broadcast, conference paper, dictionaries & encyclopedias, digital resources, infrastructure/archive/library, journal paper, linguistic resources, pre-print, repertoires and projects, report, thesis, social media, workshop.
2. Riduzione a quattro etichettature e relativa classificazione dei dati in precedenza annotati in:
    1. letteratura convenzionale:
        1. Book: (book, book paper, dictionaries & encyclopedias)
        2. Journal

20 https://arxiv.org/

21 CoRR è parte del servizio di editoria online arXiv. Quest'ultimo è proprietà della Cornell University.

3. Conference proceedings (workshop, abstract & demo)

2. letteratura non convenzionale o letteratura grigia:[22]

1. Fonti elettroniche, social media, tesi, pre-print, report, manualistica e documentazione tecnica.

In Illustrazione 11 è visibile il trend citazionale generale di CLiC-IT e AIUCD. Dal grafico affiora la distribuzione di flussi documentali che vanno a caratterizzare gli eventi in questione:

- Book e journal: AIUCD ha una prevalenza citazionale di libri e riviste;
- Conference proceeding: CLiC-it ha una prevalenza citazionale di convegni;
- Grey literature: in AIUCD la letteratura grigia viene più citata.

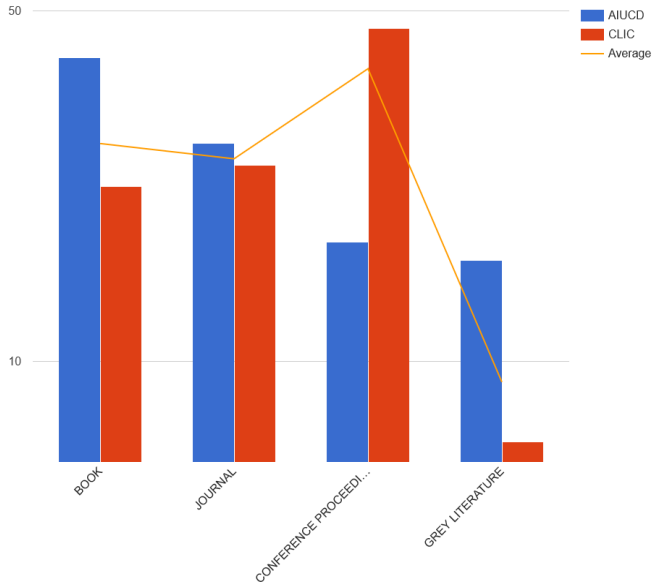

Illustrazione 11: Confronto tra fonti citazionali di AIUCD e CLiC-it.

Vediamo questi flussi documentali più nel dettaglio:

- Book e journal: in AIUCD è naturale e necessario il ricorso alla citazione di monografie multilingua (si veda Illustrazione 11) distribuite in differenti periodi storici in seno alle discipline delle scienze umane e sociali: nel quadriennio preso da

22 "Grey Literature defines an innovative approach and methodology for a wide information dissemination and exchange, by offering the web-based sharing facilities and distributed access to openly available scientific and technical document repositories, possibly under authoritative content management. An updated re-definition of GL should take into consideration the key notions of digital medium, web-based distribution channels, information access policy and access and management tools for GL. [...] At its core, Grey Literature is about producing and distributing the seeds of new knowledge" 38..

noi in considerazione, 12 monografie sono edite nel XIX secolo e 34 nella prima metà del '900; in CLiC-it è più raro il ricorso alla citazione di monografie estranee al XX e XXI secolo: nel quadriennio è stata rilevata infatti una sola citazione riferibile al XIX secolo e 19 occorrenze riferibili alla prima metà del '900. Si osserva però che la citazione più datata è riconducibile all'edizione di CLiC-it del 2017 e concerne il lessico di latino medievale del Du Cange 1678-1887. Per gli articoli in rivista la differenza tra le due aree citazionali è trascurabile. Si osserva che in AIUCD la tendenza a citare riviste in lingue europee come italiano, francese, spagnolo e tedesco è lievemente maggiore rispetto a quella di CLiC-it. In entrambe le conferenze la maggior parte delle citazioni richiama comunque contributi a riviste di lingua inglese. Per fornire un esempio, in CLiC-it nel 2017, su un totale di 297 citazioni di articoli in riviste scientifiche, soltanto 8 si riconducono a testate italiane, una rivista è francese, una è spagnola e una è catalana; le restanti 286 citazioni sono articoli di periodici scientifici in inglese.

- Conference proceedings: una caratteristica importante delle citazioni in CLiC-it rispetto a AIUCD è il maggior ricorso ad articoli in atti di convegno, distribuiti su un nucleo centrale di conferenze: *Sem/SemEval, ACL, COLING, CoNLL, EACL, EMNLP, HLT, HLT, IJCNLP, IJCNLP, LREC, NAACL, PACLIC, SIGs, TACL; ACM conference, AAAI conference, IEEE Conference; AI*IA, CLIC-it, SLI. Significativi e numerosi sono i workshops orbitanti all'interno delle voci inerenti le conferenze come ad esempio il SIGDAT-Workshop organizzato congiuntamente alla conferenza ACL. In AIUCD le voci dei convegni si intrecciano con alcune conferenze di CLiC-it, TEI, ACM, IEEE, AAAI, LREC, sebbene siano a livello disciplinare piuttosto diversificate, come esemplificato in Tabella 7.

- Grey literature: l'analisi citazionale delle voci rileva non pochi riferimenti etichettabili come letteratura non convenzionale. In alcuni casi si tratta di nuove tipologie di comunicazione scientifica attraverso l'uso di social media. In CLiC-it come in AIUCD il ricorso a citazioni di materiale presente in rete è costante sia per la consultazione in linea di archivi digitalizzati, sia per la consultazione di risorse e strumenti linguistici: e.g., corpora, linee guide, lessici, thesauri, strumenti di Information Retrieval, Machine Translation, Word Sense Disambiguation, motori di ricerca. Tuttavia, per questa tipologia documentaria esistono differenze non trascurabili tra i due eventi: (i) nelle bibliografie dei contributi AIUCD gli autori indicano riferimenti bibliografici ai social media; (ii) al contrario, le bibliografie degli articoli CLiC-it non citano i social media come tipologia di comunicazione scientifica ma come oggetto di ricerca in articoli, riviste, libri e convegni;[23] (iii) le linkografie in AIUCD sono abbondantemente usate;[24] (iv) all'interno della letteratura grigia

23 Ad esempio: "Mark Dredze Michael J. Paul. 2011. You are what you tweet: Analyzing twitter for public health. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, pages 265–272".

24 Esempi relativi alla conservazione del patrimonio digitale: "Library of Congress, PREMIS –

rientrano anche *e-print*, recuperati numerosi in CLiC-it.

Pratesi, A. 1977. «Limiti e difficoltà dell'uso dell'informatica per lo studio della forma diplomatica e giuridica dei documenti medievali». In Informatique et histoire médiévale. Atti del convegno di studi, Roma, 20-22 maggio 1975, a cura di L. Fossier, A. Vauchez e C. Violante, 187–190. Roma: École Française de Rome.
Irigoing, J., e G. P. Zarri, cur. 1979. La Pratique des ordinateurs dans la critique des textes: Paris, [Colloque international], 29-31 mars 1978. Paris.
Brincken, D. von den. 1986. «Inter spinas principum terrenorum. Annotazioni sulle summe e sui compendi storici dei Mendicanti». In Aspetti della letteratura latina del secolo XIII . Atti del primo Convegno internazionale di studi dell'AMUL, Perugia 3-5 ottobre 1983, a cura di C. Leonardi e G. Orlandi, 77–103. Firenze-Perugia.
Brown, P. F. 1991. «Aligning sentences in parallel corpora». In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, 169–176. Berkeley.
Chiesa, P., e L. Pinelli, cur. 1994. Gli autografi medievali, problemi paleografici e filologici. Atti del convegno di studio della Fondazione Elio Franceschini, Erice, 25 settembre-2 ottobre 1990. Premessa di C. Leonardi (37–60). Firenze.

Tabella 7: Esempi di citazioni ad atti di conferenze in AIUCD.

L'ultimo studio fatto sulle citazioni ha riguardato la lingua usata nelle voci bibliografiche.

Roberto Busa S.J. e Antonio Zampolli iniziarono il loro intervento al *Colloque international sur la mécanisation et l'automation des recherches linguistiques* nel 1966 a Praga con questa descrizione: "Le domaine de nos recherches comprend 9 langages en 4 alphabets aussi bien: les langues latine, italienne, allemande et anglaise en alphabet latin: les langues hébraïque, araméenne et nabathéenne en alphabet hébraïque, la langue grecque en alphabet grec et récemment la langue russe en alphabet cyrillique" 39.. Anche le bibliografie delle due conferenze italiane, benché in gran parte in inglese, come da Illustrazione 12, contengono non poche voci a lingue appartenenti a differenti alfabeti come l'italiano, il francese, il tedesco, lo spagnolo, il latino, l'arabo, il russo, il croato, il polacco, il rumeno, il sardo e il catalano. Molte di queste voci si riferiscono a documentazione cartacea come libri, riviste, atti di convegni e tesi di laurea e di dottorato. Alcune voci fanno riferimento a istituzioni nazionali coinvolte negli studi linguistici, il cui materiale è consultabile in rete come negli esempi a seguire:

- Istituto per la lingua tedesca: portale per la lessicografia scientifica;
- Scuola Normale Superiore: Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS);

Preservation Metadata: Implementation Strategies, v. 3.0 (http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf)"; OCLC. PREMIS (PREservation Metadata: Implementation Strategies) Working Group, 2005 (http://www.oclc.org/research/projects/pmwg/).

- UNED: Repertorio métrico digital de la poesía medieval castellana.

Per la lingua latina, le voci si riferiscono soprattutto a lessici e a collezioni di testi digitalizzati presenti in rete, come il seguente riferimento recuperato in AIUCD 2017: "*Analecta Hymnica Digitalia and Analecta carmine medii aevii*".

Spunti di riflessioni circa queste voci bibliografiche analizzate nel paragrafo potrebbero emergere dalla tendenza della citazione ai social media e al materiale consultabile in rete in AIUCD, e dalla tendenza alla citazione di articoli in atti di convegni in CLiC-it.

La vocazione al minimalismo nel redigere una bibliografia, come ad esempio l'indicare solo il link a un social, può far immaginare una popolazione più giovane di ricercatori.

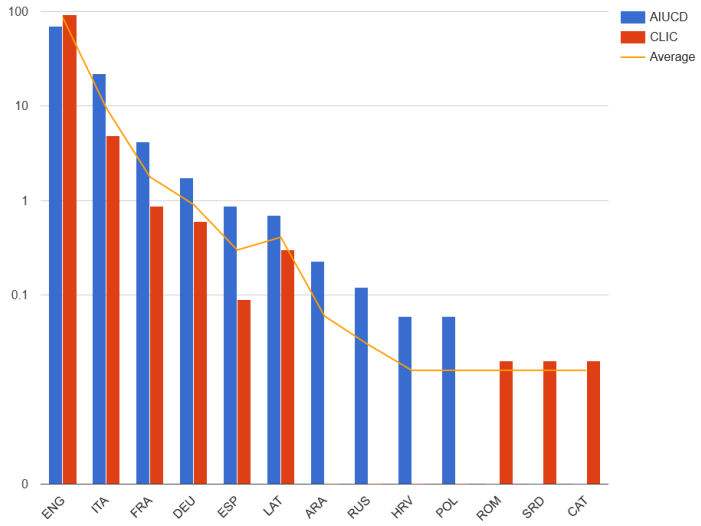

Illustrazione 12: Distribuzione delle lingue delle voci citate in AIUCD e CLiC-it

Conclusioni e Sviluppi Futuri

Le analisi descritte in questo articolo hanno preso in considerazione diverse dimensioni allo scopo di fornire uno studio esauriente delle recente ricerca italiana in due campi che storicamente hanno sorgenti comuni, ovvero le Digital Humanities e la Linguistica Computazionale. La relazione tra questi ambiti, che sono in continua espansione sia a livello nazionale che internazionale, era stata oggetto di indagine nel numero speciale "NLP and

Digital Humanities" dell'Italian Journal of Computational Linguistics (IJCoL) 88. In quel caso però gli articoli raccolti nella rivista affrontavano la questione solo indirettamente, mostrando ognuno uno specifico esempio del rapporto tra i due campi di ricerca. In questo lavoro, invece, gli oggetti dell'indagine sono stati i contributi presentati in 4 anni di conferenze AIUCD e CLiC-it: il fatto di avere a disposizione più anni di contributi ha permesso di adottare anche una prospettiva diacronica, utile per evidenziare alcuni andamenti longitudinali.

I contributi sono stati analizzati nelle loro varie parti con metodi e strumenti specifici:

1) L'analisi della collaborazione tra autori si è focalizzata sui metadati di ciascun contributo: l'applicazione di tecniche mutuate dall'analisi delle reti sociali ha evidenziato una maggiore frammentazione della comunità AIUCD rispetto a quella di CLiC-it. Un aumento delle collaborazioni tra istituzioni diverse e tra Italia ed estero è quindi auspicabile perché porterebbe ad una crescita della coesione all'interno delle DH italiane. Al contrario, AIUCD è più bilanciata rispetto a CLiC-it in termini di rappresentanza di genere anche se la diminuzione di donne autrici di contributi nelle ultime due edizioni potrebbe essere un campanello d'allarme per l'inclusività, un concetto fondamentale nelle DH a livello mondiale. Anche il calo di partecipazione da parte di aziende in AIUCD è una tendenza da invertire: la presenza del mondo non accademico alle conferenze aggiunge un particolare punto di vista, più applicativo, alla comunità e permette ai giovani ricercatori di entrare in contatto con una diversa realtà del mondo del lavoro.

2) L'analisi del contenuto si è basata sul processamento dei titoli e dei testi dei contributi attraverso l'uso di strumenti e script automatici. Identificare nuovi termini introdotti di anno in anno nei contributi delle due conferenze ha rivelato un interesse per temi e tecnologie differenti o una diversa tempistica della loro adozione da parte delle comunità: si è notato ad esempio che alcuni approcci della linguistica computazionale vengono adottati molto prima in CLiC-it e solo successivamente in AIUCD. L'estrazione di concetti-chiave ha evidenziato, tra le altre cose, la centralità dell'asse sintagmatico in AIUCD e dell'asse paradigmatico in CLiC-it. La successiva identificazione di concetti-chiave comuni ed il tracciamento del trend di rilevanza di alcuni di essi nel tempo ha posto l'attenzione verso un gruppo di temi comuni come quello di *corpus* e di aspetto *semantico*. Abbiamo poi notato che i contesti d'uso degli stessi concetti-chiave possono essere molto diversi tra i contributi delle due conferenze e che tali contributi sono stilisticamente e statisticamente differenti in quanto a vocabolario usato.

3) L'analisi citazionale ha avuto come unità di indagine le voci bibliografiche inserite dagli autori nell'articolo a testimoniare le fonti della propria ricerca. Questa analisi ha reso evidente l'impegno di molte associazioni ad operare come editori di atti di convegni e workshop. L'ampia diversità di associazioni citate è indice di multidisciplinarità in entrambe le conferenze. Gli atti di convegni e workshop sono però una fonte citazionale tipica di CLiC-it ma meno presente in AIUCD in cui, la peculiare componente storico-filologica, porta gli autori a citare più libri anche con

una dimensione temporale ben più estesa di quella di CLiC-it. Un'altra differenza importante riguarda il riferimento alla letteratura grigia. Benché in entrambe le conferenze ci siano numerose citazioni di materiale presente in rete, in CLiC-it mancano alcune categorie documentali emergenti in AIUCD come i social media, mentre sono molto presenti gli e-print. Infine l'inglese è la lingua più frequente sia nelle voci bibliografiche di CLiC-it che di AIUCD. Le citazioni in italiano sono però più frequenti in AIUCD rispetto a CLiC-it. In aggiunta, entrambe le conferenze contengono anche voci appartenenti a varie altre lingue ed alfabeti, come russo e arabo.

Sulla base delle conclusioni sopra riportate, pensiamo che potrebbe essere positivo per AIUCD incentivare la partecipazione delle aziende proponendo all'interno della conferenza delle tavole rotonde con rappresentanti delle imprese che si occupano di DH. Questo tipo di attività, organizzata già in CLiC-it 2015 e 2016, permetterebbe ai giovani ricercatori di entrare in contatto con una diversa realtà del mondo del lavoro. Inoltre, per rinforzare la connessione tra DH e linguistica computazionale, potrebbe essere utile organizzare tutorial o workshop specificamente ad essa dedicati.

Per quanto riguarda i possibili lavori futuri, diverse direzioni possono essere intraprese.

Le tecniche di analisi delle reti sociali, ad esempio, potrebbero essere applicate sia all'analisi del contenuto che a quella citazionale. Nel primo caso, si andrebbero a creare delle reti di co-occorrenza dei concetti-chiave per evidenziare le relazioni tra di essi ed identificare cluster e quindi domini di interesse nelle comunità. Tradizionalmente le reti di co-occorrenza vengono create a partire dalle keyword scelte dagli autori 53.: usare i concetti-chiave estratti automaticamente dal testo costituirebbe quindi una innovazione che potenzialmente potrebbe portare a far emergere temi diversi e meno evidenti. Un passo in questa direzione è stato fatto sviluppando un metodo e rilasciando uno script che, prendendo in input l'output di KD, crea liste di archi da importare in Gephi 63.. Per quanto riguarda l'analisi citazionale, sarebbe interessante costruire reti di citazioni e co-citazioni che, soprattutto se sviluppate in maniera diacronica, aiuterebbero ad evidenziare i processi di scoperta e consolidamento tipici dello sviluppo della conoscenza scientifica 54.. Un altro aspetto importante da investigare in futuro è quello della diffusione della citazione di materiali ad accesso aperto al fine di valutare l'impatto dell'Open Access nel campo delle DH e della linguistica computazionale 55..

In generale, la metodologia usata in questo articolo può essere estesa ad altre comunità e produzioni scientifiche del settore. Ad esempio sarebbe interessante confrontare la conferenza italiana sulle DH con quelle organizzate in altri paesi europei, come DH Benelux o Digital Humanities in the Nordic Countries, e CLiC-it con CLIN (Computational Linguistics in the Netherlands) o la conferenza TALN/RECITAL in Francia. Tali confronti aiuterebbero a capire come la comunità italiana si configura nel panorama europeo.

Acknowledgments

Gli autori vogliono ringraziare Silvia Giannini per la sua collaborazione nella stesura dell'abstract presentato ad AIUCD 2017 40. da cui è partito il lavoro descritto in questo articolo.

References

1. Manning, Christopher, Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven and David McClosky. 2014. "The Stanford CoreNLP natural language processing toolkit." In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55-60. Baltimore: Association for Computational Linguistics.
2. Aprosio, Alessio Palmero and Giovanni Moretti. 2017. "Italy goes to Stanford: a collection of CoreNLP modules for Italian." *arXiv* 1609.06204. http://arxiv.org/abs/1609.06204.
3. Chiu, Wen Ta, and Yuh Shan Ho. 2007. "Bibliometric analysis of tsunami research." *Scientometrics* 73 (1): 3–17. https://doi.org/10.1007/s11192-005-1523-1.
4. Su, Hsin Ning, and Pei Chun Lee. 2010. "Mapping knowledge structure by keyword co-occurrence: A first look at journal papers in Technology Foresight." *Scientometrics* 85 (1): 65–79. https://doi.org/10.1007/s11192-010-0259-8.
5. Ciotti, Fabio. 2017. "What's in a Topic Model? Critica teorica di un metodo computazionale per l'analisi del testo." *Testo e Senso* 18:1-11.
6. Blei, David M. 2012. "Probabilistic topic models." *Communications of the ACM* 55 (4): 77-84.
7. Hall, David, Jurafsky, Dan, and Christopher D Manning. 2008. "Studying the History of Ideas Using Topic Models." In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, October 2008*, 363–71. Honolulu: Association for Computational Linguistics. https://doi.org/10.3115/1613715.1613763.
8. Yau, Chyi Kwei, Porter, Alan, Newman, Nils and Arho Suominen. 2014. "Clustering scientific documents with topic modeling." *Scientometrics* 100 (3):767–86. https://doi.org/10.1007/s11192-014-1321-8.
9. Moretti, Giovanni, Sprugnoli, Rachele and Sara Tonelli. 2015. "Digging in the Dirt: Extracting Keyphrases from Texts with KD." In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, edited by C. Bosco, S. Tonelli e F. M. Zanzotto, 198–203. Torino: Accademia University Press.

https://doi.org/10.4000/books.aaccademia.1518.

10. Ackoff, Russell L. 1989. "From data to wisdom." *Journal of Applied Systems Analysis* 16 (1): 3–9. https://doi.org/citeulike-article-id:6930744.

11. Lenci, Alessandro, Montemagni, Simonetta and Vito Pirrelli. 2005. *Testo e computer. Introduzione alla linguistica computazionale*. Roma: Carocci.

12. Ciula, Arianna, and Cristina Marras. 2016. "Circling around texts and language: towards pragmatic modelling in Digital Humanities." *DHQ: Digital Humanities Quarterly* 10 (3). http://www.digitalhumanities.org/dhq/vol/10/3/000258/000258.html

13. Ciula, Arianna, and Øyvind Eide. 2017. "Modelling in digital humanities: Signs in context." In Digital Scholarship in the Humanities, 32 Suppl. 1. https://doi.org/10.1093/llc/fqw045.

14. Eder, Maciej, Rybicki, Jan and Mike Kestemont. 2016. "Stylometry with R: A Package for Computational Text Analysis." *The R Journal* 16 (1):1–15.

15. Craig, Hugh, and Arthur F. Kinney. 2009. *Shakespeare, computers, and the mystery of authorship*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511605437.

16. Burrows, John. 2007. "All the way through: Testing for authorship in different frequency strata." *Literary and Linguistic Computing* 22 (1) : 27–47. https://doi.org/10.1093/llc/fqi067.

17. Schreibman, Susan, and Ann M. Hanlon. 2010. "Determining Value for Digital Humanities Tools." *Digital Humanities Quarterly* 4 (2). http://www.digitalhumanities.org/dhq/vol/4/2/000083/000083.html

18. Warwick, Claire, Terras, Melissa, Galina, Isabel, Huntington, Paul, and Nikoleta Pappa. 2007. "Evaluating digital humanities resources: The LAIRAH project checklist and the internet Shakespeare editions project. In *Openness in digital publishing : awareness, discovery, and access : proceedings of the 11th International Conference on Electronic Publishing, Vienna, June 13-15, 2007*. 297-306. Vienna: ELPUB.

19. Nussbaumer, Alexander, Steiner, Christina, Hillemann, Eva-Catherine and Dietrich Albert. 2016. "User Interface Design and Evaluation in the Context of Digital Humanities and Decision Support Systems." Paper presented at *Digital Scholarly Editions as Interfaces*, Graz, Austria.

20. Anthony, Laurence. 2004. "AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit." In *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*, 7-13. Waseda University. https://core.ac.uk/download/pdf/144458559.pdf

21. Wasserman, Stanley, and Katherine Faust. 1994. *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press. https://doi.org/10.1525/ae.1997.24.1.219.

22. Melin, Göran, and Olle Persson. 1996. "Studying research collaboration using co-authorships." *Scientometrics* 36 (3): 363–77.

23. Liu, Xiaoming, Bollen, Johan, Nelson, Michael L., and Herbert Van de Sompel. 2005. "Co-authorship Networks in the Digital Library Research Community." *Information Processing & Management* 41 (6):1462–80. https://doi.org/10.1016/j.ipm.2005.03.012.

24. Bastian, Mathieu, Heymann, Sebastien and Mathieu Jacomy. 2009. "Gephi: An Open Source Software for Exploring and Manipulating Networks." In *Third International AAAI Conference on Weblogs and Social Media*, 361–62. Menlo Park, CA:AAAI Press. https://doi.org/10.1136/qshc.2004.010033.

25. Catherine Hill, Corbett, Christianne, and Andresse St. Rose. 2010. *Why So Few? Women in Science, Technology, Engineering, and Mathematics*. Washington: American Association of University Women. https://doi.org/10.1002/sce.21007.

26. Bordalejo, Barbara. 2016. "Diversity in Digital Humanities." In *Proceedings of DH Benelux*.

27. Vogel, Adam, and Dan Jurafsky. 2012. "He Said, She Said: Gender in the ACL Anthology." In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, edited by R. E. Blanche, 33–41. http://web.stanford.edu/~jurafsky/vogeljurafsky12.pdf.

28. Wernimont, Jacqueline. 2015. "Introduction to Feminisms and DH special issue." *Digital Humanities Quarterly* 9 (2).

29. Weingart, Scott. 2014. "Acceptances to Digital Humanities 2015." http://www.scottbot.net/HIAL/?p=41041

30. Posner, Miriam. 2015. "Humanities Data: A Necessary Contradiction." Paper presented at the Harvard Purdue Data Management Symposium on June 17, 2015, in Cambridge, Massachusetts. http://miriamposner.com/blog/humanities-data-a-necessary-contradiction/.

31. Kleinberg, Jon, and David Easley. *Network, Crowds, and Markets: Reasoning about a Highly Connected World. Journal of Chemical Information and Modeling*. Cambridge: Cambridge University Press, 2013. https://doi.org/10.1017/CBO9781107415324.004.

32. Grandjean, Martin. 2016. "A social network analysis of Twitter: Mapping the digital humanities community." *Cogent Arts & Humanities* 3 (1). https://doi.org/10.1080/23311983.2016.1171458.

33. Moody, James, e James Moody. 2004. "The Structure of a Social Science

Collaboration Network: Disciplinary Cohesion from 1963 to 1999." *American Sociological Review* 69 (2):213–38. https://doi.org/10.1177/000312240406900204.

34. Newman, Mark E. J. 2001. "The structure of scientific collaboration networks." In *Proceedings of the national academy of sciences* 98 (2):404–09. https://doi.org/10.1073/pnas.98.2.404

35. Lagemann, Ellen Condliffe. 1989. "The Plural Worlds of Educational Research." *History of Education Quarterly* 29 (2):183–214. https://doi.org/10.2307/368309.

36. Buitelaar, Paul, and Thomas Eigner. 2009. "Expertise mining from scientific literature." In *Proceedings of the fifth international conference on Knowledge capture - K-CAP '09*, 171-2. New York: ACM. https://doi.org/10.1145/1597735.1597767

37. Banks, Marcus. 2010. "Blog posts and tweets: The next frontier for grey literature." In *Grey Literature in Library and Information Studies*, edited by Dominic J. Farace and Joachim Schöpfel, 1-8. Berlin: De Gruyter Saur. https://doi.org/10.1515/9783598441493.

38. Marzi, Claudia, Pardelli, Gabriella and Manuela Sassi. 2011. "A terminology based re-definition of Grey Literature." In *Twelfth International Conference on Grey Literature: Transparency in Grey Literature, Grey Tech Approaches to High Tech Issues*, 19–23. Amsterdam: TextRelease.

39. Busa, Roberto S. J., and Antonio Zampolli. 1968. "Centre pour l'automation de l'analyse linguistique (C.A.A.L.), Gallarate." In *Les Machines dans la Linguistique*, Prague: Èditions de l'Académie Tchécoslovaque des Sciences.

40. Pardelli, Gabriella, Giannini, Silvia, Boschetti, Federico, and Riccardo Del Gratta. 2017. "AIUCD e CLiC-it: citazioni bibliografiche a confronto." In *AIUCD 2017 Conference*, 38–50. Roma: Università Sapienza.

41. Bird, Steven, Dale, Robert, Dorr, Bonnie J., Gibson, Bryan, Joseph, Mark T. Min-Yen Kan, Lee, Dongwon, Powley, Brett, Radev, Dragomir, and Yee Fan Tan. 2008. "The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics." In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, 1755-9. Marrakech: ELRA.

42. Dunaiski, Marcel, Greene, Gillian J. and Bernd Fischer. 2017. "Exploratory search of academic publication and citation data using interactive tag cloud visualizations." *Scientometrics* 110 (3):1539–71. https://doi.org/10.1007/s11192-016-2236-3.

43. Hassan, Saeed Ul, Safder, Iqra, Akram, Anam and Faisal Kamiran. 2018. "A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis." *Scientometrics* 116 (2):973–96. https://doi.org/10.1007/s11192-018-2767-x.

44. Johri, Nikhil, Ramage, Daniel, McFarland, Daniel and Daniel Jurafsky. 2011. "A

Study of Academic Collaboration in Computational Linguistics with Latent Mixtures of Authors." In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 124–32. Portland, OR: Association for Computational Linguistics.

45. Liu, Xiaoming, Bollen, Johan, Nelson, Michael L. and Herbert de Sompel. 2005. "Co-authorship Networks in the Digital Library Research Community." In *Information Processing & Management* 41 (6):1462–80. https://doi.org/10.1016/j.ipm.2005.03.012

46. Mariani, Joseph, Paroubek, Patrick, Francopoulo, Gil and Olivier Hamon. 2014. "Rediscovering 15 Years of Discoveries in Language Resources and Evaluation: The LREC Anthology Analysis." In *Lrec 2014 - Ninth International Conference on Language Resources and Evaluation*, edited by N. Calzolari et alii, 4632–69.

47. Munster, Sander, and Marinos Ioannides. 2015. "A scientific community of digital heritage in time and space." In *2015 Digital Heritage*, edited by I.E.E.E., 267–74. https://doi.org/10.1109/DigitalHeritage.2015.7419507.

48. Tang, Muh Chyun, Yun Jen Cheng, and Kuang Hua Chen. 2017. "A longitudinal study of intellectual cohesion in digital humanities using bibliometric analyses." *Scientometrics* 113 (2):985–1008. https://doi.org/10.1007/s11192-017-2496-6.

49. Taşkın, Zehra, and Umut Al. 2017. "A content-based citation analysis study based on text categorization." *Scientometrics* 114 (1):335–57. https://doi.org/10.1007/s11192-017-2560-2.

50. Thompson, Jennifer Wolfe. 2002. "The death of the scholarly monograph in the humanities? Citation patterns in literary scholarship." *Libri* 52 (3):121–36. https://doi.org/10.1515/LIBR.2002.121.

51. White, Howard D., and Katherine W. McCain. 1988. "Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972–1995». *Society* 49 (4):327–55.

52. Nerbonne, John, and Sara Tonelli. 2016. "Introduction to the Special Issue on Digital Humanities of the Italian Journal of Computational Linguistics." *Italian Journal of Computational Linguistics* 2 (2): 7–10.

53. Radhakrishnan, Srinivasan, Erbis, Serkan, Isaacs, Jacqueline A. and Sagar Kamarthi. 2017. "Correction: Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature." PLoS ONE 12 (9):185771. https://doi.org/10.1371/journal.pone.0185771.

54. Yan, Erjia, and Cassidy R. Sugimoto.2011. "Institutional interactions: Exploring social, cognitive, and geographic relationships between institutions as demonstrated through citation networks." *Journal of the American Society for Information Science and Technology* 62 (8):1498–1514. https://doi.org/10.1002/asi.21556

55. Di Donato, Francesca. 2010. "Le sfide dell'Open Access al sistema di comunicazione della scienza." *La Rivista SIFP* 2.

56. Fogel, Ephim G. 1965. "The Humanist and the Computer: Vision and Actuality." In *American Behavioral Scientist*, edited by J. B. Bessinger, S. M. Parrish, e H. F. Arader, 37–40. New York: IBM Corporation. https://doi.org/10.1177/000276426500900407.

57. Del Gratta, Riccardo, Goggi, Sara, Pardelli, Gabriella and Nicoletta Calzolari. 2018. "LREMap, a Song of Resources and Evaluation." In *Proceedings of LREC*, edited by N. Calzolari et alii, 1275-1281.

58. Bartolini, Roberto, Goggi, Sara, Monachini, Monica and Gabriella Pardelli. 2018. "The LREC Workshops Map." In *Proceedings of LREC*, edited by N. Calzolari et alii, 557-562.

59. Radev, Dragomir, Muthukrishnan, Pradeep and Vahed Qazvinian. 2009. "The ACL Anthology Network." *Language Resources and Evaluation* 47:54–61.

60. Wang, Qing. 2018. "Distribution features and intellectual structures of digital humanities: A bibliometric analysis." *Journal of Documentation* 74 (1):223–46. https://doi.org/10.1108/JD-05-2017-0076.

61. Terras, Melissa. 2006. "Disciplined: Using educational studies to analyse "Humanities Computing"." In *Literary and Linguistic Computing*, 21:229–46. https://doi.org/10.1093/llc/fql022.

62. Montanari, Elio. 2003. *La critica del testo secondo Paul Maas: testo e commento*. Firenze: Sismel - Edizioni del Galluzzo.

63. Sprugnoli, Rachele and Giovanni Moretti. 2019. "Discovering Research Themes in Scientific Research: from Keyphrase Extraction to Co-occurrence Networks". Abstract. Book of Abstract of AIUCD 2019, forthcoming.